Stata 软件基本操作和数据分析入门

第三讲 概率分布和抽样分布

概率分布累积函数

1. 标准正态分布累积函数 **norm(X)**
2. **t** 分布右侧累积函数 **ttail(df，X)** ，其中 **df** 是自由度
3. χ^2 分布累积函数 **chi2(df，X)** ，其中 **df** 是自由度
4. χ^2 分布右侧累积函数 **chi2tail(df，X)** ，其中 **df** 是自由度
5. **F** 分布累积函数 **F(df1，df2，X)**，**df1** 为分子自由度，**df2** 为分母自由度
6. **F** 分布右侧累积函数 **F(df1，df2，X)**，**df1** 为分子自由度，**df2** 为分母自由度

累积函数的计算使用

正态分布计算

X 服从 **N(0,1)**，计算概率 **P(X<1.96)**

```
. display norm(1.96)
.9750021
```

即概率 **P(X<1.96)＝0.9750021**

display 可简写为 **di**，如： **di norm(1.96)**，同样可以得到上述结果。

X 服从 **N(0,1)**，计算概率 **P(X>1.96)**，则

```
. di  1- norm(1.96)
.0249979
```

即概率 **P(X>1.96)＝0.0249979**

X 服从 **N(μ,σ²)**，则 $Y=\frac{X-\mu}{\sigma}\sim N(0,1)$，因此对其他正态分布只要在函数括号中插入一个上述表达式就可以得到相应概率。

例如：X 服从 $N(100,6^2)$，计算概率 P(X<111.76)，则操作如下

```
. di    norm((111.76-100)/6)

.9750021          即：概率 P(X<111.76)=0.9750021
```

又如 X 服从 $N(100,6^2)$，计算概率 P(X>90)，操作如下

```
. di 1-norm((90-100)/6)

.95220965
```

χ^2 分布累积概率计算

设 X 服从自由度为 1 的χ^2 分布，计算概率 P(X>3.84)，则操作如下

```
. di    1-chi2(1,3.84)

.05004353                  概率 P(X>3.84)=0.05004353
```

设 X 服从自由度为 3 的χ^2 分布，计算概率 P(X<5)，则操作如下

```
. di    chi2(3,5)

.82820288          概率 P(X<5)=0.82820288
```

χ^2 分布右侧累积概率计算

设 X 服从自由度为 1 的χ^2 分布，计算概率 P(X>3.84)，则操作如下

```
. di chi2tail(1,3.84)

.05004353                  概率 P(X>3.84)=0.05004353
```

设 X 服从自由度为 3 的χ^2 分布，计算概率 P(X<5)，则操作如下

```
.di    chi2(3,5)

.82820288          概率 P(X<5)=0.82820288
```

t 分布右侧累积概率计算

设 t 服从自由度为 10 的 t 分布，计算概率 P(t>2.2)，操作如下

```
. di ttail(10,2.2)

.02622053       概率 P(t>2.2)=0.02622053   (注意：这是右累积函数)
```

设 t 服从自由度为 10 的 t 分布，计算概率 P(t<－2)，操作如下

```
. di 1-ttail(10,-2)

.03669402       概率 P(t<－2)=0.03669402
```

F 分布累积概率计算

设 F 服从 F(3,27)，计算概率 P(F<1)，操作如下：

```
. di F(3,27,1)          注意这里的函数是大写 F，stata 软件中是区分大小写的

.59208514           概率 P(F<1)=0.59208514
```

设 F 服从 F(4,40)，计算概率 P(F>3)，操作如下：

```
. di    1-F(4,40,3)

.02954694           概率 P(F>3)=0 .02954694
```

F 分布右侧累积概率计算

设 F 服从 F(3,27)，计算概率 P(F<1)，操作如下：

```
. di    1-Ftail(3,27,1)          注意这里的函数是大写 F，stata 软件中是区分大小写的

.59208514           概率 P(F<1)=0.59208514
```

设 F 服从 F(4,40)，计算概率 P(F>3)，操作如下：

```
. di    Ftail(4,40,3)

.02954694           概率 P(F>3)=0 .02954694
```

概率分布的临界值计算

正态分布的临界值计算函数 **invnorm(P)**

例如：双侧 $U_{0.05}$(即：左侧累积概率为 **0.975**)，操作如下

```
. di invnorm(0.975)
1.959964        即 U0.05＝1.959964
```

t 分布的临界值计算函数 **invttail(df,P)**

例如计算自由度为 **28** 的右侧累积概率为 **0.025** 的临界值 $t_{28,\alpha}$，操作如下

```
. di invttail(28,0.025)
2.0484071               临界值 t28, α=2.0484071
```

χ^2 分布的临界值计算函数 **invchi2(df,P)** 或 **invchi2tail(df,P)**

例如：计算自由度为 **1** 的χ^2右侧累积概率为 **0.05** 的临界值$\chi^2_{0.05}$，操作如下：

```
. di invchi2(1,0.95)
3.8414591               临界值χ2 0.05=3.8414591
```

或者操作如下：

```
. di invchi2tail(1,0.05)
3.8414591               临界值χ2 0.05=3.8414591
```

F 分布的临界值计算函数 invF(df1,df2,P) 或 invF(df1,df2,P)

例如计算分子自由度为 3 和分母自由度 27 的右侧累积概率为 0.05 的临界值，操作如下：

```
. di invF(3,27,0.95)
2.9603513
```

临界值 $F_{0.05}(3,27)= 2.9603513$

或者操作为：

```
. di invFtail(3,27,0.05)
2.9603513
```

临界值 $F_{0.05}(3,27)= 2.9603513$

产生随机数

计算机所产生的随机数是通过一串很长的序列数模拟随机数，故称为伪随机数，在实际应用这些随机数时，这些随机数一般都能具有真实随机数的所有概率性质和统计性质，因此可以产生许许多多的序列伪随机数，一个序列的第一个随机数对应一个数，这个数称为种子数(seed)，因此可以利用种子数，使随机数重复实现。

设置种子数的命令为 **set seed** 数。每次设置同一种子数，则产生的随机序列是相同的。

产生(0,1)区间上的均匀分布的随机数 **uniform()**

例如产生种子数为 100 的 20 个在(0,1)区间上的均匀分布的随机数，则操作如下：

```
clear                清除内存
set seed 100         设置种子数为 100
```

set obs 20　　设置样本量为 20

gen r=uniform()　产生 20 个在(0，1)区间上均匀分布的随机数。

list　　显示这些随机数

结果如下

	r
1.	.7185296
2.	.1646728
3.	.9258041
4.	.1833736
5.	.0067327
6.	.7413361
7.	.3599943
8.	.1634543
9.	.445553
10.	.6489049
11.	.3799431
12.	.5964895
13.	.0251346
14.	.2164402
15.	.6848479
16.	.1270018
17.	.6466258
18.	.1869288
19.	.4522384
20.	.067132

利用均匀分布随机数进行随机分组：

例：某实验要把 20 只大鼠随机分为 2 组，每组 10 只，请制定随机分组方案和措施。

第一步、把 20 只大鼠编号，1，2，3，4，5，6，7，8，9，10，11，12，13，14，15，16，17，18，19，20。并且标明。

第二步、用 Stata 软件制定随机分组方案，操作如下：

clear　　清除内存

set seed 200	设置种子数为 **200**
set obs 20	设置样本量为 **20**
range no 1 20	建立编号 **1** 至 **20**
gen r=uniform()	产生在**(0,1)**均匀分布的随机数
gen group=1	设置分组变量 **group** 的初始值为 **1**
sort r	对随机数从小到大排序
replace group=2 in 11/20	设置最大的 **10** 个随机数所对应的记录为第 **2** 组，即：最小的 **10** 个随机数所对应的记录为第 **1** 组
sort no	按照编号排序
list	显示随机分组的结果

结果如下：

	no	r	group
1.	1	.9512007	2
2.	2	.5249876	2
3.	3	.5129986	1
4.	4	.126439	1
5.	5	.5866161	2
6.	6	.7059209	2
7.	7	.2633286	1
8.	8	.5644688	2
9.	9	.1171033	1
10.	10	.954065	2
11.	11	.4822863	1
12.	12	.3347736	1
13.	13	.5678902	2
14.	14	.7994431	2
15.	15	.1180503	1
16.	16	.9834299	2
17.	17	.2807874	1
18.	**18**	**.095245**	**1**
19.	**19**	**.9446051**	**2**
20.	**20**	**.3467524**	**1**

随机分组整理如下

第一组										
编号	3	4	7	9	11	12	15	17	18	20
第二组										
编号	1	2	5	6	8	10	13	14	16	19

产生服从正态分布 **N(μ，σ^2)**的随机数 **invnorm(uniform())*σ+μ**。例如产生 **10** 个服从正态分布 **N(100，6^2)**的随机数，操作如下:

clear	清除内存
set seed 200	设置种子数为 **200**
set obs 10	设置样本量为 **10**
gen x=invnorm(uniform())*6+100	产生服从 **N(100，6^2)**的随机数
list	显示随机数

结果如下：

	x
1.	109.9397
2.	100.3761
3.	100.1955
4.	93.13968
5.	101.3131
6.	103.249
7.	96.2013
8.	100.9739
9.	92.86244
10.	110.1137

教学应用：考察样本均数的分布。

由于个体变异的原因，样本均数 $\bar{X}$ 的抽样误差(其定义为样本均数与总体均数的差值)是不可避免的，并且样本均数的抽样误差是呈随

机变化的。对于一次抽样而言，无法考察样本均数的抽样误差的规律性，但当大量地重复抽样，计算每次抽样的样本均数 $\bar{X}$，考察样本均数 $\bar{X}$ 的随机分布规律性和统计特征。举例如下：

利用计算机模拟产生 **100000** 个服从正态分布 **N(100,6^2)**的样本，样本量分别为 **n=4**，**n=9**，**n=16**，**n=36**，每个样本计算样本均数。这里关键处是要清楚什么是样本量(每次抽样所观察的对象个数，也就是每个样本的个体数 **n)**、什么是样本个数(指抽样的次数)，现以 **n=4** 为例，一条记录存放一个样本，样本量 **n=4**，也就是每个样本的第 **1** 个数据放在第 **1** 列，第 **2** 个数据放在第 **2** 列，第 **3** 个数据放在第 **3** 列，第 **4** 个数据放在第 **4** 列，因此第 **1** 行是第一个样本，第 **2** 行是第 **2** 个样本，第 **100000** 行是第 **100000** 个样本，计算样本均数放在第 **5** 列，因此共有 **100000** 个样本均数。具体操作如下：

clear	清除内存
set memory 60m	扩大虚拟内存为 **60M**
set obs 100000	设置记录数为 **100000**
set seed 200	设置种子数为 **200**
gen x1=invnorm(uniform())*6+100	产生第 **1** 个随机数据
gen x2=invnorm(uniform())*6+100	产生第 **2** 个随机数据
gen x3=invnorm(uniform())*6+100	产生第 **3** 个随机数据
gen x4=invnorm(uniform())*6+100	产生第 **4** 个随机数据
gen mean=(x1+x2+x3+x4)/4	计算平均数，并且存放在变量名为 **mean**
su mean	以样本均数为数据，计算其平均值和标准差

结果

Variable	Obs	Mean	Std. Dev.	Min	Max
mean	100000	99.98388	3.002225	87.97424	112.0461

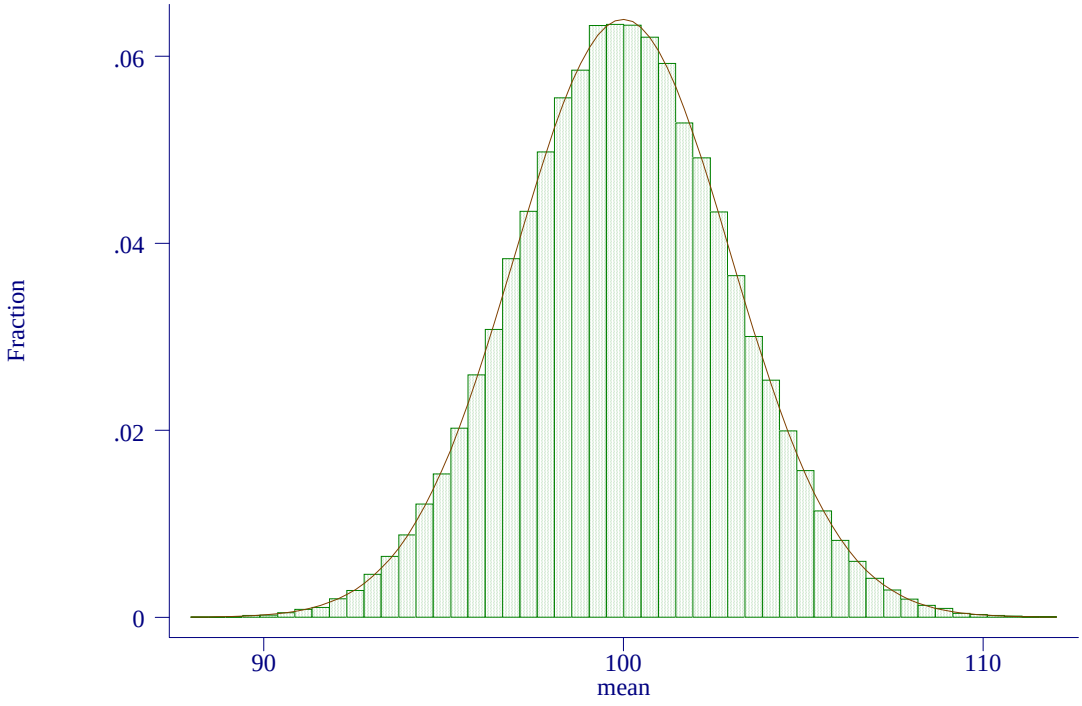
现共有 **100000** 个样本，每个样本计算一个样本均数，因此有 **100000** 个样本均数，现在把一个样本均数 $\bar{x}$ 视为一个数据，把 **100000** 个样本均数视为一个样本量为 **100000** 的新样本(这个样本里有 **100000** 个 $\bar{x}$)，计算这 **100000** 个 $\bar{x}$ 的平均值和标准差：得到：

这 **100000** 个 $\bar{x}$ 的平均值＝99.98388 非常接近总体均数μ=100

这 **100000** 个 $\bar{x}$ 的标准差＝3.002225$\approx \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{4}} = 3$ (理论上可以证明样本均数的总体均数与样本所在的总体的总体均数相同，样本均数的标准差=$\frac{\text{样本所在总体的总体标准差}}{\sqrt{n}}$)

再考察这 **100000** 个 $\bar{x}$ 的频数图

graph mean,bin(50) xlabel ylabel norm



可以发现正态分布的样本均数仍呈正态分布，峰的位置在μ＝100。

再考察这 **100000** 个 $\bar{X}$ 的百分位数

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf.	Interval]
mean	100000	2.5	94.11224	94.05934	94.15675
		5	95.04831	95.00758	95.08677
		50	99.97672	99.95568	100.0002
		95	104.9248	104.8881	104.9571
		97.5	105.8656	105.8161	105.9181

比较理论上的百分位数

百分位数	Stata 操作	理论百分位数	模拟百分位数
$P_{2.5}$	di 100+invnorm(0.025)*3	94.120108	94.11224
P_{5}	di 100+invnorm(0.05)*3	95.065439	95.04831
P_{50}	di 100+invnorm(0.5)*3	100	99.97672
P_{95}	di 100+invnorm(0.95)*3	104.93456	104.9248
$P_{97.5}$	di 100+invnorm(0.975)*3	105.87989	105.8656

可以发现理论上的百分位数与模拟数据的百分位数非常接近。可以证明：样本量越大，这种 $\bar{X}$ 的误差小的可能性越大。

由于在实际研究中，只有一个样本，因此只有一个样本均数，无法如模拟数据一样计算样本均数的标准差，但是一个样本的数据可以计算样本的标准差 **S** 近似σ，利用样本均数的标准差 $\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}$ 关系，间接估计得到样本均数的标准差估计为 $S_{\bar{X}}=\frac{S}{\sqrt{n}}$，为了区分样本的标准差和样本均数的标准差，故称 $S_{\bar{X}}=\frac{S}{\sqrt{n}}$ 为标准误。

为了帮助大家方便地进行模拟实习，特地编制的相应的 **stata** 模拟程序:模拟正态分布的样本均数分布的模拟程序 **simumean.ado** 复制到 **stata** 软件安装的目录下的子目录 **ado\base**。例如：**stata** 软件安装在

D:\stata，则 simumean.ado 复制到 d:\stata\ado\base

然后启动 **stata** 软件后，输入连接命令:**net set ado d:\stata\ado\base**

若 **stata** 安装在其他目录下，则相应改变上述路径便是(这是一次性操作，以后无需再重复进行)。这是模拟抽 **10000** 个正态分布的样本，具体说明如下：

举例说明

simumean　样本量　均数 标准差

例如模拟抽 **10000** 个正态分布的样本，样本量为 **4**、总体均数是 **20**、标准差为 **6**，则操作如下：

simumean　4　20　6

得到下列结果(随机的)

Variable	Obs	Mean	Std. Dev.	Min	Max
mean	10000	19.99352	2.990616	8.344506	31.40937
ssd	10000	5.511469	2.346368	.258496	15.51934

即 10000 个样本均数(视为一个新的样本数据)的平均值为 19.99352≈总体均数 20，10000 个样本均数的标准差＝$2.990616 \approx \frac{6}{\sqrt{4}} = \frac{\text{总体标准差}}{\sqrt{n}} = 3$。

变量　样本量　%　百分位数

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf.	Interval]
mean	10000	2.5	14.19629	14.01392	14.31436
		5	15.08899	14.96281	15.2017
		50	19.96537	19.88963	20.03251
		95	24.91111	24.78268	25.05202
		97.5	25.92742	25.75092	26.05995

理论上，样本均数 $\bar{X}$ 的 95%范围是$\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$ ＝20±1.96×3=(14.12,25.88)

比较 **10000** 个样本均数的 **95**％百分位数=**(14.196,25.927)**

模拟习题

1)运行正态分布的样本均数模拟程序 simumean.ado，考察不同样本量情况下，$\bar{X}$ 的标准差与$\frac{\sigma}{\sqrt{n}}$的差异，95％范围的比较。

样本量 n	9	16	25	36	49
总体均数μ	100	100	100	100	100
总体标准差σ	6	6	6	6	6
$\bar{X}$ 的标准差					
$\frac{\sigma}{\sqrt{n}}$					
$\mu\pm1.96\frac{\sigma}{\sqrt{n}}$					
$P_{2.5}-P_{97.5}$					

考察频数图的变化

graph 变量名,xlabel bin(40)

考察原始资料：**graph x1,xlabel bin(40)**

考察样本均数(变量名为 **mean**) **graph mean,xlabel bin(40)**

考察：原始资料和样本均数的峰的位置，离散程度。

考察非正态分布情况下，样本均数

可以运行下列程序

双峰分布的样本均数分布程序：**simubpeak.ado**

自由度为 **1** 的χ^2分布的样本均数模拟程序 **simuchi.ado**

把上述程序复制到 路径:**\stata\ado\base**

连接：**net set ado** 路径:**\stata\ado\base**

操作：**simubpeak.ado** 样本量

simuchi.ado 样本量

考察原始资料的分布和样本均数的分布变化，

原始资料所在总体分布的频数图：**graph x1,bin(40) xlabel**

样本均数的抽样分布的频数图：**graph meanx ,bin(40) xlabel**

考察原始资料 **x1,x2** 的标准差和样本均数 **meanx** 的标准差

样本量 n	9	16	25	36	100

考察不同样本量对样本均数分布的影响。

可以证明：样本量较大时，样本均数的分布趋向于正态分布(称为中心极限定理)，并且样本均数的总体均数(理论均数)仍与样本所在总体相同，样本均数的总体标准差(标准误)＝$\frac{\text{样本所在总体的总体标准差}\sigma}{\sqrt{n}}$